6.3 Regresja logistyczna

Zaimplementuj uczenie regresji logistycznej metodą **gradient descent**(4). Uczenie powinno obsługiwać zarówno zatrzymanie po osiągnięciu zadanego błędu jak i wykonaniu zadanej ilości iteracji. Wytrenuj klasyfikator na **syntetycznych jednomodowych zbiorach** danych i porównaj jego działanie z sklearn.linear_model.LogisticRegression.

$$\Delta\vec{w_j} = -\eta(t_j - y_j)f(\vec{x_j}^T w_{\vec{j-1}})[1 - f(\vec{x_j}^T w_{\vec{j-1}})]\vec{x_j} \qquad \text{gdzie } f(s) = \frac{1}{1+e^{-\beta s}} \qquad (4)$$

- $\Delta\vec{w_j} \in \mathbb{R}^D$ j-ta korekta wektora wag
- $\eta \in \mathbb{R}_{>0}$ krok algorytmu uczenia (learning rate)
- $t_j \in \mathbb{R}$ etykieta j-tej próbki ze zbioru uczącego
- $y_j \in \mathbb{R}$ predykcja modelu dla j-tej próbki ze zbioru uczącego
- $\vec{x_j} \in \mathbb{R}^D$ j-ta próbka ze zbioru uczącego
- $w_{\vec{j-1}} \in \mathbb{R}^D$ wagi modelu po $(j-1)$-tej (poprzedniej) korekcie

Wykorzystując zbiór Rain in Australia (© Commonwealth of Australia 2010, Bureau of Meteorology) dokonaj preprocessingu danych. Usuń kolumny mające więcej in 30% brakujących wartości (oraz kolumnę 'Risk-MM' jeżeli istnieje) i odseparuj kolumnę z wartością przewidywaną ('RainTomorrow'). Dokonaj imputacji brakujących wartości zakładając że są one MCAR (Missing Completely At Random) - zmienne kategoryczne należy zastąpić dominantą a dane numeryczne medianą. Dokonaj winsoryzacji danych odstających ponad 1.5 IQR. Podziel dane dla każdego z regionów na zbiory testowe i treningowe ze stratyfikacją. Znormalizuj dane numeryczne i zakoduj (one-hot) dane kategoryczne. Należy zwrócić szczególną uwagę na cykliczny charakter komponentów w daty - data jest dyskretną numeryczną wartością złożoną. Kodowanie kolumny 'Location' nie jest konieczne. Należy także zwrócić uwagę na potencjalny wyciek danych. Naucz osobny model LogisticRegression dla każdego z regionów. Który model ma najwyższą skuteczność? Porównaj skuteczność najlepszego modelu ze skutecznością własnej implementacji uczonej na tym samym zbiorze.

Sprawdź skuteczność modeli regionalnych na krajowym zbiorze testowym zbudowanym ze zbiorów testowych wszystkich regionów. Który model osiągnął najwyższą skuteczność? Czy był to model o najwyższej skuteczności lokalnej? Porównaj skuteczność najlepszego modelu krajowego z sklearn.dummy.DummyClassifier. Co o prawdziwej skuteczności modelu mówi to porównanie? Czy budowanie modelu w oparciu o fragmentaryczne, lokalne dane to dobry sposób zmniejszenia wymaganej ilości danych uczących?

6.3.1 Kryteria oceniania

Porównanie sklearn i własnej implementacji klasyfikatora na syntetycznych zbiorach → 3
Poprawny preprocessing danych, nauczenie modeli dla regionów i porównanie ich skuteczności → 4
Weryfikacja skuteczności modeli regionalnych na zbiorze krajowym i porównanie skuteczności najlepszego modelu → 5